Reviewer #2

#RC2.1:In this paper, the authors propose a geostatistical framework which relies on a high-resolution regional climate model (RCM) to provide information on the precipitation climatology as well as on the anisotropy of the departures of observations from that climatology. The focus of the paper is on documenting, in particular through two case studies, the added value of the RCM for estimating the anisotropy of the covariogram. A statistical analysis of the performance of the method for 786 events is also presented. The experiment is well designed, and allows the authors to assess separately the impact of using an anisotropic covariogram and the impact of estimating the anisotropy using the RCM. A comparison against a reference interpolation method (SPAZM) is also proposed.

We thank the reviewer for this positive feedback.

#RC2.2: The introduction reads well but fails to mention a relevant paper by Khedhaouiria et al. (2022) published in NPG which proposed a method based on an ensemble of NWP models to estimate the anisotropy of innovations for optimal interpolation of precipitation.

Thank you for pointing out this missing article. We add L38: “Numerical weather prediction ensembles have also been explored (Khedhaouiria et al., 2022) to infer background error covariances in data assimilation approaches”.

#RC2.3: For the section on domain and data, I suggest including a subsection on the study period. The study period is mentioned in the section on COMEPHORE, but it would be simpler to add a section dedicated to the study period after the section on meteorological data, because the study period is constrained by the availability of COMEPHORE and AROME. A subsection on observed data should also be added. The rain gauge network is currently described in the study domain subsection.

We agree with the reviewer that more subsections are needed in the section on domain and data. We propose to move L69-73 into a separate subsection named “Rain gauge observations”. We also include a fourth subsection called “study period” with the following text: “The study period ranges from 1982 to 2018, which corresponds to the availability of AROME simulations. 786 precipitation events (nearly 20 events per year), defined as the days with at least 50 mm recorded at a minimum of five rain gauges, are selected”.

#RC2.4: In the sub-section describing the CP-RCM AROME, I would like the authors to provide more details on the model configuration, more specifically w.r.t. to the ability of the system to represent specific events and not only the climatology of precipitation over the domain. This is important since AROME is used later to inform the interpolation method on the anisotropy of the covariogram, but not on the amount of precipitation associated with the event. Figure 4.6 and 4.8 show that there are significant discrepancies between COMEPHORE and AROME precipitation fields. Is this happening because CP-RCM AROME is not sufficiently constrained by ERA-Interim or is it inherent to the predictibility of precipitation events in this region? Would we expect a similar degree of agreement for a short-term forecast of precipitation based on AROME? Is ERA-Interim only used as boundary conditions or is some form of spectral nudging used to prevent RCM model drift? How far is the study domain from the ALADIN and AROME boundaries? Given the model configuration, do we expect AROME to only be able to provide information on the precipitation climatology but yet be able to provide useful information on the anisotropy of the precipitation structures? Why would that be the case?

We agree with the reviewer that more details are needed on AROME configuration. We modify the AROME subsection:

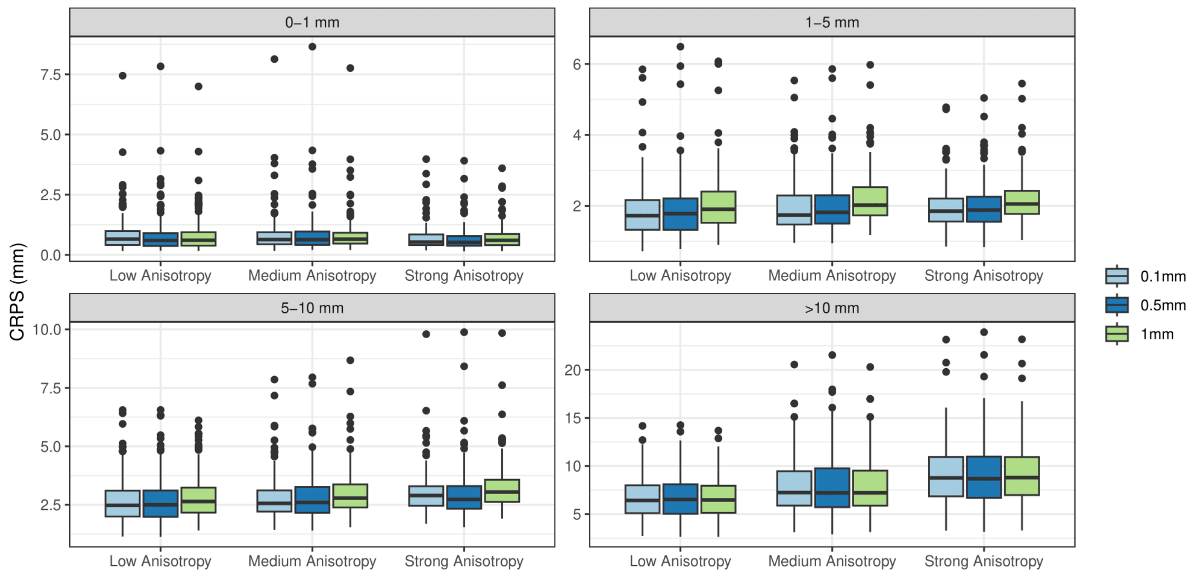
“AROME simulations (Caillaud et al., 2021) are produced with the convection-permitting RCM AROME in its NWP configuration cycle 41t1, which uses 60 vertical levels from 10 m to 1 hPa, including 21 levels below 2000 m to better resolve the lower-tropospheric dynamics over complex Alpine terrain. In this CP-RCM configuration, deep convection is explicitly resolved, while only shallow convection remains parameterized. The AROME domain over the Alpine region lies approximately 300–400 km from the lateral boundaries, which are forced by hourly outputs from the CNRM-ALADIN RCM (Nabat et al., 2020). ALADIN uses 91 vertical levels together with spectral nudging to ensure consistency with the large-scale circulation imposed by the ERA-Interim reanalysis (Dee et al., 2011). AROME simulations are available at the hourly timescale for the Alpine region, as described in the Flagship Pilot Study of the Coordinated Regional Climate Downscaling Experiment (CORDEX-FPS,Fantini et al. (2018)), at 2.5 km spatial resolution, and cover the 1982–2018 years. Hourly outputs are aggregated to a daily timescale.

Previous studies (Ban et al., 2021; Caillaud et al., 2021; Monteiro et al., 2022) show that AROME provides a more realistic representation of intense precipitation than its driving model ALADIN, despite persistent biases. In a Lagrangian evaluation over the Mediterranean region, which includes our study domain, Caillaud et al. (2021) report that AROME simulations reproduce well the location, intensity, frequency, and interannual variability of heavy precipitation events. Remaining biases are mostly due to the model, rather than insufficient constraint from ERA-Interim. In the AROME model, very intense daily amounts (> 200 mm day^{-1}) tend to be underestimated, the spatial extent of intense convective cells is overestimated, and their propagation speed is slightly too high. These biases could be reduced by further refining horizontal and vertical resolution and by improving the parameterization of residual shallow and dry convection.

Unlike a short-term NWP forecast, AROME-climate simulations do not assimilate observations such as radar reflectivity; therefore, they are not expected to reproduce individual events exactly, but rather their typical spatial structures. Consequently, even if absolute precipitation amounts may be biased, the spatial organization and anisotropy of intense precipitation systems, key for informing the anisotropic covariogram, may be sufficiently well captured by AROME to support our interpolation framework.”

#RC2.5: In the methods section, the authors choose to consider observed precipitation of less than 0.5 mm as zeros for the purpose of normalizing the precipitation field. How was this number chosen? Are results sensitive to this choice? The back-transformation introduces a bias, which the authors do not take into account (see Van Hyfte et al., 2023, Tellus A). Can the authors quantify the impact of ignoring this source of bias on their analysis?

We thank the reviewer for this remark. 0.5 mm is a standard choice in censoring daily precipitation (e.g Naveau et al. 2016). Results are not very sensitive to this choice.



The above figure highlights the CRPS score with the arANISO model. A too high threshold (1 mm) leads to bias for moderate precipitation (1-5 mm, 5-10 mm), and a too low threshold (0.1 mm) degrades the estimation of 0-1 mm. 0.5 mm corresponds to a good compromise.

We used a Quantile-Quantile mapping to transform and back-transform precipitation. We did not use the traditional Box Cox-transformation, that is a non linear transformation and therefore introduces a bias in the back-transformation step. The uncertainty of our transformation/back-transformation step only lies in the estimation of the gamma parameters. We discuss this idea in the subsection 5.3.

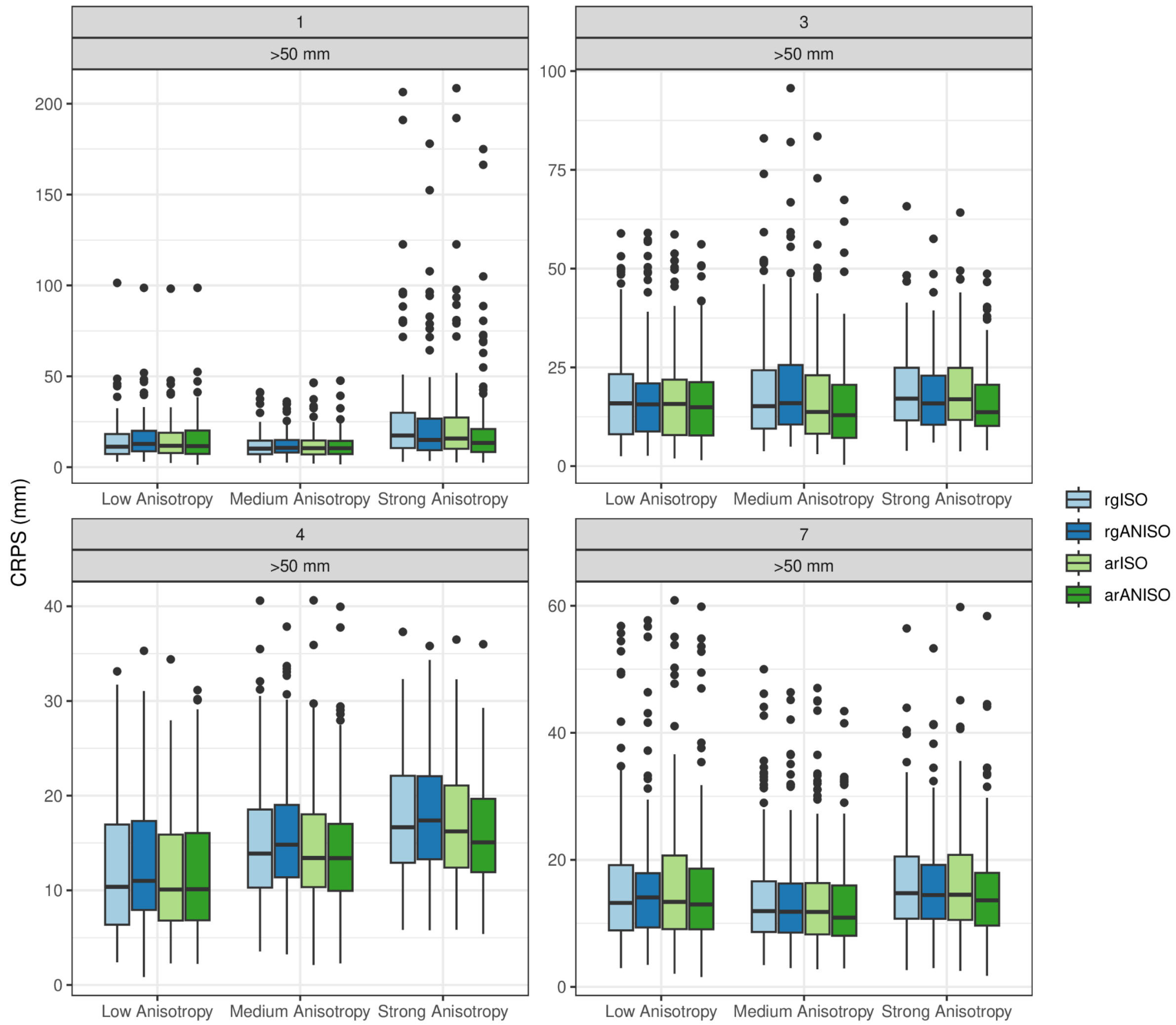
#RC2.6: The authors chose 786 precipitation events to evaluate the proposed method. It would be interesting to know more about the type of events that were selected. Please categorized them by weather regime and season. In particular, can you identify events for which orographic intensification is expected and events for which snow was observed at higher elevations? Do we expect the performance and ranking of the methods to vary depending on the type of event ?

Thank you for this suggestion. The 786 precipitation events selected in this study are predominantly associated with southerly atmospheric flow and central depression patterns, for which strong orographic intensification is typically expected over the region. During winter, some of these events also include snowfall at high elevations.

To assess whether the performance of the interpolation methods varies with event type, we stratified the 786 events according to the eight weather regimes defined in Garavaglia et al. For each regime, we computed the CRPS. Overall, the relative ranking of the methods is consistent across weather regimes. The rgANISO model provides a better covariance estimation than rgISO for regime 3 (oceanic flow). However, this improvement is not observed for regimes 4 and 7, which correspond to the heaviest precipitation events. Across all regimes, arANISO remains the best-performing model, suggesting that the superiority of this method is robust to meteorological conditions.

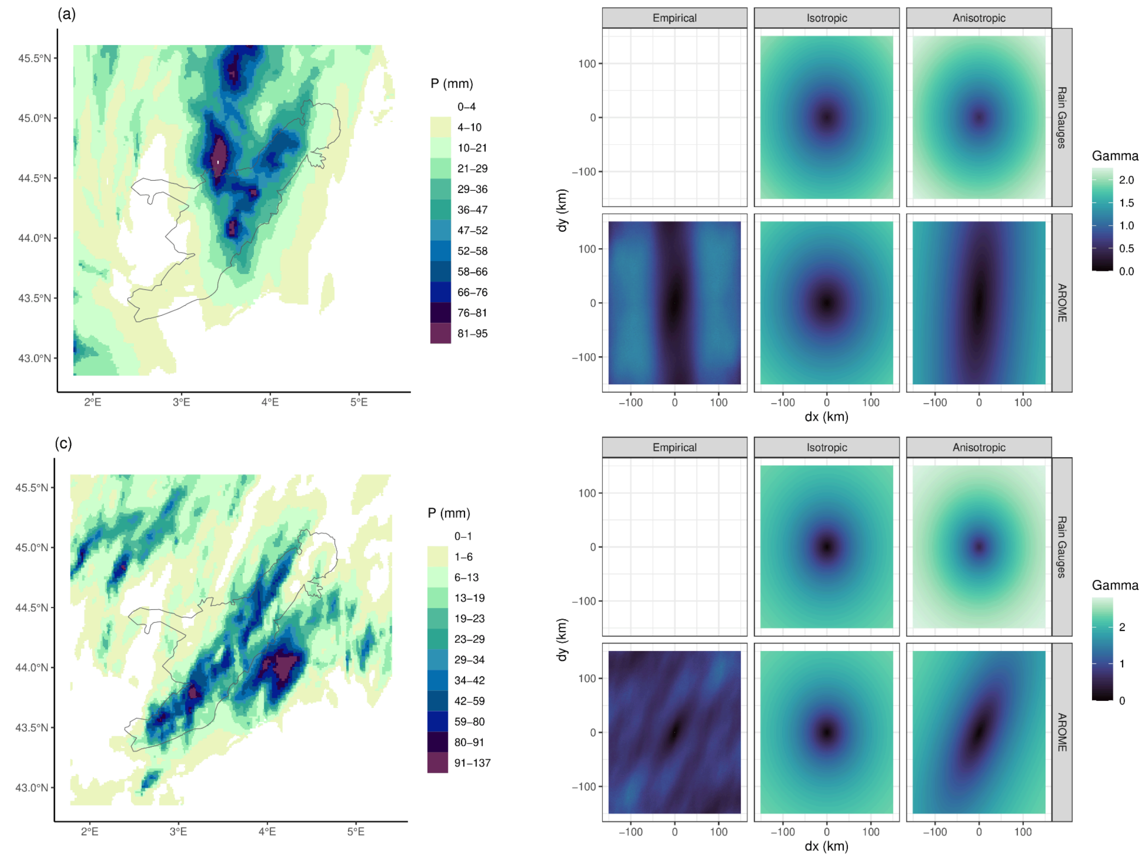
We propose to add at L70:
"Most of the 786 precipitation events arise from southerly atmospheric flow and central depression patterns, where strong orographic intensification is expected. Some winter events also include snowfall at high elevations."

We also include a figure of CRPS by weather regime in the Supplementary Material.

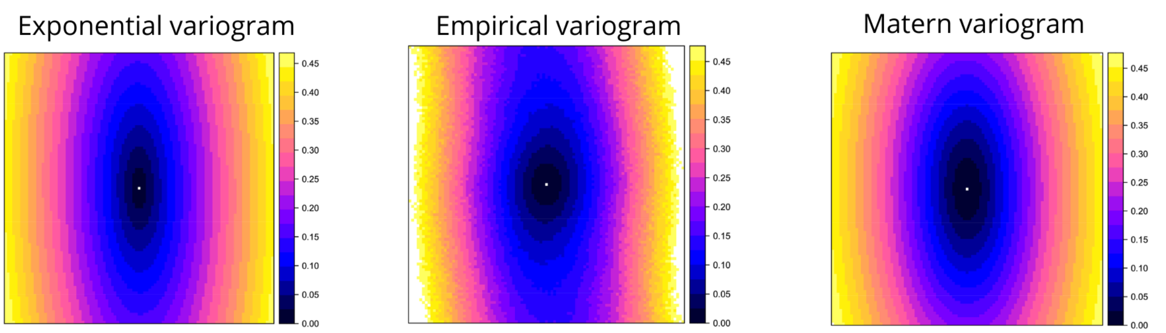


#RC2.7: In the results section, the authors should present and discuss the covariograms that are obtained for each of the two case studies. The 2D covariograms derived from AROME are presented in Figure 4.8, but that does not tell us how well it fits the experimental covariogram. Furthermore, no information is provided on the fitted covariograms for the other three experiments (rgISO, rgANISO and arISO). This is important, in particular to show that the choice of an exponential covariogram is appropriate based on the data. Did the authors check that the exponential variogram provided a good fit for the 786 events considered in this study?

We thank the reviewer for this comment. We replace Figure 8 by the below figure.



Moreover, we make visual inspections of variogram fitting.



Here are the empirical variogram and the fitted exponential and Matérn variograms for a given day. The exponential variogram fits the empirical one well, benefiting from the large number of estimation pixels used as virtual gauges. However, the Matérn variogram appears to better capture the short-range spatial variability, providing a smoother representation than the exponential variogram. A cross-validation would be needed to assess the best variogram for precipitation interpolation.

#RC2.8: In the discussion, the authors address many limitations of the method, in particular the fact that it would be difficult to apply on a larger domain. This is an important limitation, because it would seem impractical to deploy such a complex interpolation method operationally if it cannot be applied on a large domain. I encourage the authors to propose a workflow that would allow the application of the method on a larger domain. Could it not be applied watershed by watershed? Would the cost of doing so be prohibitive? Are there other solutions?

We thank the reviewer for raising this important point. We agree that the applicability of the method over a large domain is a key consideration for operational use, and that the current implementation is best suited for moderately sized regions. Extending the approach to larger and topographically complex domains indeed requires additional methodological considerations.

A first practical solution would be to apply the method watershed by watershed, or more generally to divide the study area into climatically homogeneous sub-regions. This would ensure that the covariance structure remains locally stationary while preserving hydrologically coherent areas. Simulations would remain consistent within each major watershed, which is often sufficient for hydrological applications. The main challenge is the definition of appropriate sub-region boundaries and areas, as overly small regions may lose spatial coherence while overly large regions may violate stationarity assumptions.

A more robust solution for large-scale applications would involve adopting non-stationary covariance models. Some geostatistical approaches include non-stationary covariance:
(i) geographical coordinate deformation to map complex terrain into a space where covariance is closer to stationary, or
(ii) locally stationary covariance models in which parameters evolve spatially but do so smoothly across the domain.
Such approaches would allow the method to be applied on much larger and more heterogeneous domains without the need to arbitrarily define sub-regions, but with heavier computing times.

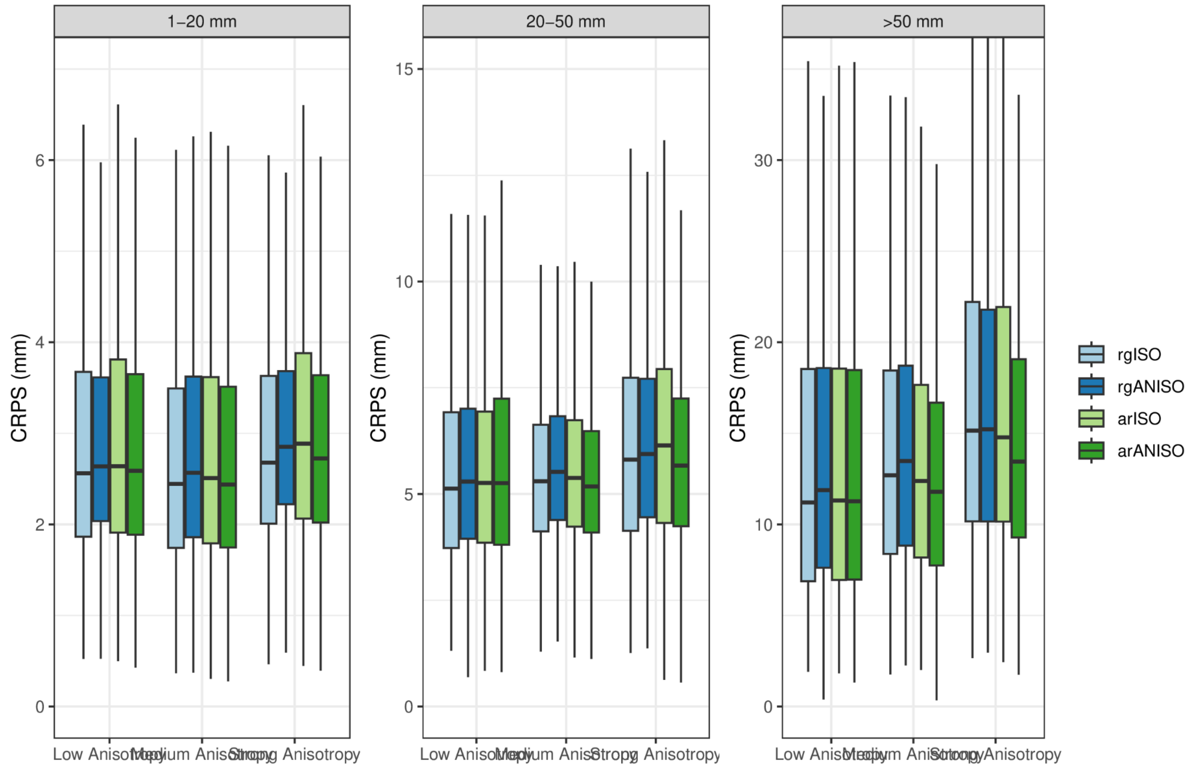
We have added the following text at L330:
"Extending the method to larger and topographically complex domains would require non-stationary covariance. A practical option is to partition the region into climatologically homogeneous sub-regions, ideally preserving major watershed boundaries to maintain hydrological consistency. Alternatively, a more scalable solution is to incorporate non-stationary covariance structures, for example, through geographical coordinate deformation (Youngman, 2023) or locally stationary covariance models (Paciorek and Schervish, 2006; Risser and Calder, 2017), which would allow spatial dependence to evolve smoothly across the domain. These approaches would make the method suitable for operational applications over larger domains."

#RC2.9: Although evaluating the method on significant events (where more than 50mm/day was observed by at least 5 gauges), in practice one would likely want to apply the method for all days. Did the authors assess how well the method performs for less intense precipitation events?

We thank the reviewer for this remark. In this study, we compare covariance estimation on a collection of 786 daily events, and compute evaluation metrics only on the stations exceeding 50 mm/day. However, we agree that in practical applications the method would need to perform well across a full range of precipitation intensities.

We provide in the below figure the metrics for the same set of 786 daily events, including additional precipitation intensity classes (1-20 mm, 20-50 mm). The ranking is similar for the three precipitation classes considered, indicating that the conclusions drawn for intense precipitation events generalize to lighter precipitation.

We also include a figure of CRPS by precipitation intensity class in the Supplementary Material.



#RC2.10: The authors also mention in the discussion the possibility of applying the method using numerical weather forecasts rather than using a RCM. I think this is worth discussing in more details. In particular, I would expect the precipitation field of a short-term forecast to correlated better with observed precipitation, and thus it might be possible to infer more from the forecast than simply the anisotropy of the precipitation field. Furthermore, one might have access to an ensemble of weather forecasts.

Thank you very much for this remark. We agree that numerical weather forecasts (NWP) should exhibit a higher correlation with observed precipitation than CP-RCM simulations, due to the assimilation of past radar reflectivity. Consequently, it is likely that NWP forecasts could provide not only the anisotropy of the precipitation field, but also information on precipitation intensity and spatial variability. Moreover, the provision of ensemble NWP forecasts would allow us to quantify additional interpolation uncertainty.

We have added at L334:
"NWP assimilate past radar reflectivity and should therefore display a higher correlation with observations than CP-RCM simulations. As a result, NWP may allow us to extract precipitation intensity, spatial patterns and spatial variability, while quantifying interpolation uncertainty through conditional simulations and the use of ensemble NWP forecasts. A natural follow-up would be to use NWP forecasts as both drift and covariance structures within a kriging-with-external-drift framework (e.g. Velasco 2009, Schiemann 2011)"

#RC2.11: Finally, one important aspect of precipitation interpolation that is not discussed in this paper is the issue of quality control. When interpolating precipitation observations, in particular in complex terrain, the issue of quality control is central because it can be very difficult to identify problematic observations based on neighboring stations, given the impact of orography on precipitation amounts over short distances. I understand that this issue might be out of scope for this paper, but I wonder if it was an issue for the authors when applying the method over 786 events. Was the observed data quality controlled? How? Could the presence of outliers impact the results of your analysis? Could your method be used to improve the quality control process through the use of cross-validation? This would likely be crucial to address before the method can be used for real-time applications.

We agree with the reviewers that quality control is central before providing gridded precipitation analysis, especially over a long period where rain gauges can change from locations and measurement devices, which can cause temporal discontinuity. Moreover, in complex terrain, strong spatial gradients limit the ability to identify outliers (sensor malfunctions) from neighboring stations.

The observed data was not quality control in this study, which is beyond the scope of this study. Because we do not work with climatological statistics, temporal discontinuity is not a major issue. However, outliers may still occur and could affect the interpolated fields. Such outliers may influence some local results, but we do not expect them to alter the conclusions of this study.

Before applying the method in an operational or real-time context, we will consider a quality control procedure, including a spatial anomaly analysis to identify outliers that exceed physically plausible differences from nearby stations under similar terrain characteristics, followed by an homogeneity test such as the Standard Normal Homogeneity Test (SNHT; Alexandersson 1986). The R package climatol encompasses those corrections.

Moreover, missing values are commonly filled using a linear regression with nearby rain gauges as predictors. Our approach could be used as a substitute of the linear regression, providing a value, and the associated uncertainty.

We propose to add L72: “The observed data was not quality control in this study. Because we do not work with climatological statistics, temporal discontinuity is not a major issue. However, outliers may still occur and could affect the interpolated fields. Such outliers may influence some local results, but we do not expect them to alter the conclusions of this study. Before applying the method in an operational or real-time context, a quality control procedure is needed, including a spatial anomaly analysis to identify outliers that exceed physically plausible differences from nearby stations under similar terrain characteristics, followed by an homogeneity test such as the Standard Normal Homogeneity Test (SNHT; Alexandersson 1986)”.